From a specialised corpus to classrooms for specific purposes

Creating the EU Word List

Réka Jablonkai
Institute of Behavioural Studies and Communication Theory
Corvinus University of Budapest
rjr@uni-corvinus.hu

Abstract

This paper reports a corpus-based analysis of the most frequent lexical items in English EU discourse. An EU word list was developed on the basis of the English EU Discourse Corpus containing slightly more than 1 million running words of official EU texts of 40 different genres that was created on the basis of the findings of a needs analysis questionnaire among EU professionals. The established EUWL comprises 513 word families, which account for 18.03% of the tokens in the EEUD Corpus. The evaluation of the EUWL found that established word list is EU-specific and it has a high text coverage in EU texts. The results of the analysis established that the EUWL can serve as a firm basis for course and materials design for English for the EU language programmes.

1 Introduction

The language use within the European Union has been the topic of studies focusing on language policy, translation and terminology issues (McArthur, 2003; Truchot, 2002; Fischer, 2007). Truchot (2002) demonstrates the rise of the use of English not only in the communication between member states, but also in internal communication within EU institutions, especially as regards written communication. English gaining more and more grounds in the EU as *lingua franca* necessitates preparing future Hungarian EU professionals for the English language use within the EU context. Issues such as the comprehensive analysis of the variety of the English language used within the EU or teaching materials for English for the EU courses need to be addressed especially in the light of the preparation for Hungary's upcoming Presidency in the European Council.

The analysis of the language use of a specific field has always been in the centre of research in English for Specific Purposes. There have been several corpus-based studies in ESP investigating small corpora representing the language use of certain subject fields. However, there have been only a handful of studies investigating English EU texts explicitly for pedagogic purposes. Tribble (2000) analysed a specific EU genre, that was, proposals for EU funds, aiming to draw conclusions on writing skills development necessary for writing such difficult texts in English. Trebits (2008; 2009a; 2009b) investigated certain lexical items like *EU* and *trade*, conjunctions and phrasal verbs in English EU texts drawing conclusions on the importance of these aspects in teaching English for the EU courses. Furthermore, in the analysis of a corpus of 200,000 words of English EU documents she found that 46.5% of the word types are not among the BNC 3000, which means, as she concludes, that a substantial number of lexical elements in English EU texts are not part of the vocabulary of an intermediate level (B1-B2) language learner.

Although both authors examined relevant genres and linguistic aspects of English EU discourse, in order to provide future EU professionals with the appropriate English language

preparation, a more comprehensive approach to English EU discourse is needed. As a consequence, the aim of the present study is to establish a word list of EU-related vocabulary, based on English EU texts published by EU institutions, that can be used as a starting point for course and materials design for English for the EU courses.

Therefore the following research questions were formulated to guide the present analysis:

(1) What are the high-frequency lexical elements in English EU discourse?

(2) What are the EU-specific lexical elements in English EU discourse?

In what follows, ESP word lists that have been established with the help of electronic corpora will be presented. Secondly, criteria and methods for the selection of relevant elements for the EU word list based on the English EU Discourse Corpus comprising official EU texts will be discussed. Thirdly, the EU word list will be presented and evaluated by testing its coverage in EU, legal and general registers and genres. In concluding, pedagogical implications and practical applications will be discussed.

1.1 Word lists in ESP

Word lists containing the basic vocabulary for learners of English were compiled as early as the first half of the 20th century (Nation, 1997). One of the most widely known and used word list has been the General Service List of English Words edited by West (GSL) (West, 1953). The aim of his list was to establish the vocabulary learners of English as a foreign language should start with. Despite the age, some errors and the fact that it had been created based on a written corpus, the GSL is still widely referred to and applied as the first most frequent 2000 words for EFL learners (Coxhead, 2000; Nation, 1997; Wang, Liang, & Ge, 2008).

With the advances of computer science and accessibility of corpora and corpus analysis software programmes, the task of creating word lists based on frequency lists of general and specialised corpora became feasible for individual researchers and teachers of ESP. Mudraya (2006), for example, created the Student Engineering Word List (SEWL) with 1260 word families based on her 2 million running word corpus of textbooks on basic engineering disciplines, such as Engineering Mechanics, Engineering Materials, Manufacturing Processes and Computer Programming. She aimed at developing a reliable English for Engineering syllabus for students in Thailand who had to study from English-language textbooks for their engineering courses at a local university. After organising the initial frequency list of more than 18,000 word types by word families, the selection of the word families forming part of the final engineering word list was carried out on the basis of the cumulative frequency of the members of the word families. The cut-off point was set at 100 occurrences, or 0.005% of the whole corpus. The first ten headwords of the SEWL with frequency information are given in Table 1 below.

N	Headword	Frequency	%
1	**use**	10,313	0.52
2	**force**	9247	0.46
3	**form**	7075	0.35
4	flow	7045	0.35
5	**pressure**	7016	0.35
6	**show (v)**	7002	0.35
7	**determine**	6896	0.34
8	**figure**/configure	6650	0.33
9	section	6404	0.32
10	**line**	5812	0.29

Table 1 The ten most frequent word families in the SEWL based on (Mudraya, 2006)

Comparing the SEWL with the GSL shows that the word list for engineering students contains many elements of the general vocabulary which are highlighted in bold in Table 1. This might be useful for courses where students do not have a sound basis in General English before specialising in the language of a subject field, but in most of the cases students of ESP already possess the basic general vocabulary. Therefore a word list that focuses on the frequent lexical elements that are specific for the given field is more useful for course and materials design.

A way word list compilers controlled for specificity was that the word families of the GSL were excluded from among the frequently occurring word families in the specialised word lists of the subject field. The first example for such a word list is the Academic Word List (AWL) (Coxhead, 2000). The aim of the list was to replace similar earlier lists (Ghadessy, 1979; Xue & Nation, 1984) that had been compiled without the help of electronic corpora and to serve as the basis for language courses for academic purposes. The corpus used for the analysis was the 3.5 million-word Academic Corpus, which contained more than 400 academic texts like journal articles, course books and laboratory manuals. The corpus was made up of four sub-corpora: arts, commerce, law, and science containing about 875,000 running words and each was subdivided into seven subject fields, for example, Education, History, Accounting, Economics, Criminal law, Rights and remedies, Biology, and Chemistry. The selection of word families was guided by the following principles: (1) ensuring **specialised occurrence** by including word families in the final AWL that are outside the GSL representing the first 2,000 most frequent English words; (2) requiring that word families represent the lexis of several academic disciplines by determining a minimum **range**, that is, a member of a word family had to have an occurrence higher than 10 in each of the main sub-corpora and had to occur 15 times or more in the 28 subject fields; (3) setting a minimum cumulative **frequency** of occurrence of a word family at higher than 100 in the Academic Corpus.

In the course of the selection process priority was given to range over frequency in order to avoid bias by longer texts and topics. It meant that word families with more members had to have a cumulative frequency of 100, whereas word families with a single member were included with a frequency less than 100. The least frequently occurring single-member word family was the word *forthcoming* with a frequency of 80. Coxhead's final Academic Word List contains 570 word families and in order to assist the sequencing of teaching it is presented in frequency-based sublists (Coxhead, 2000).

Coxhead (2000) evaluated the AWL by testing its coverage of the Academic Corpus, another corpus of academic texts and a corpus of fiction texts. Results of these analyses indicated that the AWL is a truly academic word list as it accounted for the 10.0% of all the tokens in the Academic Corpus, it covered 8.5% of the second academic corpus and its coverage of the corpus of fiction texts was only about 1.4%.

Chen and Ge (2007) investigated the text coverage of the AWL in medical research articles (RAs). They concluded that although word families in the AWL represent a high text coverage – slightly more than 10% – and dispersion in the medical RAs, only 51.2% of all word families in the AWL were frequently used in their corpus of medical RAs. Encouraged by the findings of Chen and Ge's research Wang and his colleagues (2008) established the Medical Academic Word List (MAWL) of 623 word families frequently used across various subfields of medicine.

Following Coxhead's (2000) methodology the compilation of the MAWL was based on a one-million-word corpus of medical research articles of 32 different sub-fields of medicine like Urology, Health Informatics, Gastroenterology, Surgery, etc. and the final word families were selected according to similar criteria defining (1) **specialised occurrence**, (2) **range** and (3) **frequency**. **Specialised occurrence** was understood in the same way as in Coxhead's study, that is, only word families outside the 2000 word families of the GSL were included. **Range** was defined by the minimum number of occurrences of members of word families in the 32 sub-fields at 16, that is, word families had to be applied in at least half of the sub-fields of medicine. The criterion **frequency** was set at 30 for the cumulative occurrence of word families in the whole corpus of medical research articles. Wang and colleagues (2008) argue that because their corpus is approximately a third of the corpus of Coxhead, the criterion frequency was set at the third of Coxhead's frequency requirement of 100, that is at 30 occurrences. They also applied an additional step in the selection process of the final MAWL, which was consulting two experienced professors of English for Medical Purposes, who made decisions on the inclusion or elimination of controversial word families.

The analysis of the MAWL included testing its text coverage in the corpus of medical RAs and comparing it to the AWL. The MAWL was found to cover 12.24% of the total corpus which is slightly higher than the text coverage of the AWL of academic texts. The comparison of the two word lists showed that only 342 (54.90%) of the word families of the MAWL can be found among the word families of AWL. On the basis of these results Wang and colleagues (2008) argue that different disciplinary discourses operate with their own subject-specific lexis which makes a general academic word list less valuable for individual disciplines.

1.2 The unit of analysis: the word family

A widely used unit of analysis in research into vocabulary teaching (Nation and Kennedy, 1994; Nation and Waring, 1997; Chung and Nation, 2003), defining necessary text coverage for effortless comprehension of texts (Hirsch and Nation 1992; Ward, 1999; Nation, 2001; Nation, 2006) and developing word lists for general and specific language teaching purposes (Coxhead, 2000; Nation, 2004; Mudraya, 2006; Wang et al., 2008) has been the word family. According to Bauer and Nation (1993) a word family includes a base word, its inflected forms and transparent derivations. Transparency behind the idea of a word family refers to the assumption that understanding a derived or inflected form of a word does not require extra effort from the language learner if they are already familiar with the base word or a derived form and some knowledge of the word-building processes in English. Transparency also implies that the meaning of the base word and derived forms must be closely related, for example, *hard* and *hardly* would not be included in the same word family because of the difference in their meaning. The concept is also supported by empirical evidence as research found that the word family is a psychological unit in the mental lexicon (Coxhead, 2000; Nation, 2006).

For the creation of word families, Bauer and Nation (1993) defined seven levels of inflection and affixation based on criteria including the frequency, productivity, predictability and regularity of affixes which can be used for decisions on whether a certain word form can be

included into a word family at a given level. These levels with short descriptions and examples are illustrated in Table 2 with additional affixes in italics that have been found particularly productive in the EEUD Corpus. Bauer and Nation (1993) also state that their levels are arbitrary and further affixes can be included if they are found frequent or useful in a particular field. Table 3 illustrates the concept of the word family by two examples, *ABLE* from the GSL, and *ANALYSE* from the AWL.

Levels	Description	Examples for affixes at this level	Example
Level 1	Each form is a different word	-	develop
Level 2	Inflectional suffixes	plural, 3rd person singular present tense, comparative, possessive	develops developed developing
Level 3	The most frequent and regular derivational affixes	-able, -er, -ish, -ly, -ness, non-, un-	developable undevelopable developer undeveloped
Level 4	Frequent, orthographically regular affixes	-ation, -ful, -ize, -ment	development developmental
Level 5	Regular but infrequent affixes	-age, -atory, -ling, mid-, -ship, pro-, semi-, sub-	semidevelopment
Level 6	Frequent but irregular affixes	re-, pre-, -ee, -ive	redevelop predevelopment
Level 7	Classical roots and affixes, compounds	*Euro-*, *agri-*, *ex-*	agri-development

Table 2 Levels of inflection and affixation with additional subject specific affixes (based on Bauer and Nation, 1993)

Although the concept of the word family is widely applied in vocabulary studies there are some limitations to its application. A major practical difficulty of the concept is the requirement of transparency, that is, which word forms should be recognised as belonging to a particular word family. Biber (2006) reported that he found "it extremely difficult to reliably group the [remaining] words into word families" (Biber, 2006). Another difficulty is to define at which level word families should be interpreted and if there are any additional affixes to include in the analysis of the vocabulary of a particular field. Despite these limitations the present study applied the concept of the word family for making the results comparable to earlier analyses of ESP vocabulary. Potential discrepancies between what is included in individual word families were minimised by applying the 14 word family lists created by Nation (Nation, 2006) on the basis of the BNC (Heatley, Nation, & Coxhead, 2002).

Headword	Members of the word family
ABLE (GSL)	ability, abilities, inability abler, ablest, ably, unable
ANALYSE (AWL)	analysed, analysing analyser, analysers, analyses, *analysis*, analyst, analysts analytic, analytical, analytically analyze, analyzed, analyzes, analyzing

Table 3 Examples for word families from the GSL and the AWL

2 Method

2.1 The English EU Discourse Corpus

The corpus compiled for the comprehensive analysis of lexis of English EU discourse contains 1,174,753 running words from 241 written texts representing 40 different EU genres like treaties, regulations, press releases, presidency conclusions, etc. The corpus building process was preceded by a needs analysis survey among EU professionals to gather information about the type of documents they use in their jobs (Jablonkai, 2008). The results of the questionnaire survey helped selecting relevant EU genres and texts for the corpus. As the study aimed to identify lexis associated with the EU in general and not one specific EU subject field in particular, efforts were made to balance the corpus for the different subject fields of EU activities like monetary policy, economy, agriculture, security policy, education, regional policy, etc. There were altogether 34 EU subject fields defined based on the list of EU policies available on the official website of the EU (http://europa.eu/pol/index_en.htm). Only texts published by one of the EU institutions like the Commission, the Parliament, the Council, etc. were included in the corpus. The sample EU texts were kept at their original length, but the reference sections where different pieces of EU legislation are listed were deleted. Concerning the date of issue of the texts in the corpus, the majority (94%) were published in the time period between 2000-2008, 5% of the texts in the corpus were issued in the 1990s and there are a few texts (1%) from the 1980s. On the whole, the English EU Discourse Corpus represents a general view of English EU discourse that was found suitable for identifying the necessary vocabulary for courses of English for the EU.

2.2 Developing the English EU word list

The corpus analysis programmes Range (Heatley, Nation, & Coxhead, 2002) and Wordsmith Tools 4 (Scott, 2004) were used to develop the EU word list. First a frequency word list was created by Wordsmith 4 and it was organised by word families using the lemmatiser function of the programme which joins certain entries according to a pre-prepared list. This list was prepared on the basis of the 14 base word lists of the Range programme. For the fine tuning of the word list the Range programme was used. The programme counts the frequency of word types in several different files and records the frequency of occurrence of individual word types in total and in each file. It also counts the number of files in which each word type occurs. Table 4 shows the output of the Range programme for a few examples from the EEUD Corpus. The programme can also be used with different word lists of word families and it can count the cumulative frequency of a word family and provides information on the percentage of tokens, word types and number of families of a word list in a corpus (Nation, 2001). The Range software was also used to evaluate the final EU word list for text coverage in texts representing different registers and genres.

Word type	Range	Total frequency	Frequency in File 1	Frequency in File 2	Frequency in File 3
Community	34	3222	80	69	185
framework	34	1010	19	1	50
implement	32	227	12	2	4
OJ	32	874	25	28	55
undertakings	24	254	7	0	2

Table 4 Sample output of range and frequency by the Range programme

The three selection criteria used by Coxhead (2000) were adopted in this study with some modifications. Firstly, **specialised occurrence** was ensured by eliminating the word families of the GSL from among the word types of the frequency list of the EEUD Corpus. Secondly, only word families used in a wide range of EU topics were selected. In her study, Coxhead used a two-level criterion for defining range for the word selection. As her Academic Corpus was divided into sub-corpora of disciplines and these were further divided into subject areas, she set a range for both levels for a member of a word family. On the other hand, Wang and colleagues (2008) set the criterion for range at 50%, that is, word families with a cumulative range of 16 or more of the total 32 sub-fields of medicine, were included in the final Medical Academic Word List. As the aim of the EU Word List was to provide a list of useful words for students of EU studies with an intermediate level proficiency of general English, the criteria for word selection were set slightly broader than in previous studies. Word families had to occur in 16 or more of the 34 EU related subject fields which correspond to a **range** of 47%. Thirdly, this study as that of Wang and colleagues (2008) started out from the cumulative **frequency** criterion set by Coxhead (2000) at 100 in her 3.5 million word corpus. Wang et al., however, argued that because their corpus of medical RAs had 1 million running words, which is approximately a third of that of Coxhead's corpus, they set the frequency criterion at 30 for inclusion into their word list assuming a linear relationship between the number of running words and the number of word types in a corpus. Biber (2006), however, based on experiments of the stability of vocabulary distribution, found that the relationship between corpus size and the number of word types is not linear. According to his findings half a corpus represents around 70% of the word types in the larger corpus. The simple formula suggested for adjusting the number of word types in corpora of different sizes is:

# of word types of Corpus 1 = # of word types of Corpus 2 / square root of corpus size
(Biber, 2006:256)

The simple formula was reformulated to calculate the number of word types of Corpus 2:

# of word types in Corpus 2 = number of word types in Corpus 1 × Square root of relative corpus size of Corpus 2 to Corpus 1

Thus a corpus (Corpus 2) of half the size of another corpus (Corpus 1) has a number of word types of 0.707 times the number of word types of the full corpus as the square root of 0.5 is 0.707. Applying the formula to the required cumulative frequency of 100 applied by Coxhead to her three times bigger corpus than the corpus used in the present study, resulted in the adjusted cumulative frequency of 57 for inclusion into the EU word list as the square root of one-third is 0.57.

On the basis of these considerations the final selection criteria were formulated as follows:

1. **Specialised occurrence**: The word families included in the final EU Word List had to be outside the GSL representing the first 2,000 most frequent English words.
2. **Range**: A member of a word family had to occur 16 or more times in the 34 EU related subject fields.
3. **Frequency**: The cumulative frequency of occurrence of a word family had to be higher than 57 in the English EU Discourse Corpus.

In addition to these criteria two experts were also consulted on finalising the EU word list. One of them was an EU expert and the other was an experienced teacher of ESP. Involving experts was found necessary as earlier studies on ESP lexis also report on the difficulties of identifying subject-specific technical and semi-technical vocabulary (Chung & Nation, 2004; Mudraya, 2006) and consulting ESP experts in the final stage of developing a word list (Wang et al., 2008; Bowker & Pearson, 2002)

In the selection process range was considered secondary to frequency because all the texts in the corpus were issued by EU institutions and as such by definition all represented an EU related subject field. Therefore, following the consultation with the two experts another 28 word families were added to the final list which met the first and third criteria fully. These were selected despite the fact that their range was less than 16 because their cumulative frequency was high and were considered necessary for language learners for an EU context by the experts. Thus the word families that can be found in the least of the EU related subject fields are *ICT, interoperability, democracy* and *statutory* with a range of 12 and a cumulative frequency of 166, 75, 103 and 272 respectively.

Evaluation of the final EUWL was carried out with the help of the Range programme (Heatley et al., 2002) by testing the text coverage of the list in several registers and genres of different sources. These included another corpus of English EU texts (Trebits, 2008), randomly selected pieces of EU legislation and EU press releases, randomly selected news texts with business, UK news, world news and Europe news topics, news releases of the UK government, two randomly selected pieces of UK legislation, and British and American literary texts. The 20th century literary texts were downloaded from the Project Guthenberg's collection of texts and the extract from Dickens's Tale of two cities came as a trial text with the Wordsmith Tools (Scott, 2004).

3 Results

3.1 Elements of the EU word list

The final EUWL contains 513 word families that are made up of 2,457 words. Table 5 gives two examples for the word families with its members in the EUWL. A list of the one hundred most frequent word families of the EU word list can be found in Appendix 1. The word families among the most frequent ones include *EUROPE, COMMISSION, COMMUNITY, REGULATION, FINANCE* and *IMPLEMENT* and as can be seen in Table 6 almost all of them occurred in all the EU-related subject fields. Examples for the least frequent word families are *CAMPAIGN, VULNERABLE, WORLDWIDE, HIGHLIGHT* and *ALIGN* and as Table 7 shows all of them occurred at least 57 times in the corpus and their range is at least 40% which means that words with the lowest frequency occurred in at least 14 EU-related subject fields.

N	headword	Cumulative frequency	%	Members of the word family
1	EUROPE	7401	0.69%	europe[600] europe's[90] cross-europe[1] e-europe[11] european[6621] european-based[1] european-wide[1] europeans[29] intra-european[3] non-european[20] trans-european[23] transeuropean [1]
3	COMMUNITY	3635	0.34%	community [3113] communities[355] community's[109] community-based[6] community-wide[10] community-flagged[6] extra-community[1] intra-community[20] non-community[15]

Table 6 Examples for EU word families

N	EU Word family	Cumulative family frequency	Range	% of whole corpus
1	*EUROPE*	7401	100.00%	0.69%
2	*COMMISSION*	5390	100.00%	0.50%
3	*COMMUNITY*	3635	100.00%	0.34%
4	*REGULATE*	2693	97.06%	0.25%
5	*FINANCE*	2693	100.00%	0.25%
6	*IMPLEMENT*	2285	100.00%	0.21%
7	*PROCEED*	2229	100.00%	0.21%
8	*EC*	2172	100.00%	0.20%
9	*TREATY*	1996	100.00%	0.19%
10	*POLICY*	1925	100.00%	0.18%
11	*EU*	1883	94.12%	0.17%
12	*REQUIRE*	1869	100.00%	0.17%
13	*AUTHORITY*	1864	100.00%	0.17%
14	*ESTABLISH*	1766	100.00%	0.16%
15	*DIRECTIVE*	1753	85.29%	0.16%

Table 6 The first fifteen word families of the EUWL

N	EU Word family	Cumulative family frequency	Range	% of whole corpus
499	*DERIVE*	64	67.65%	0.01%
500	*COMPULSORY*	63	52.94%	0.01%
501	*EEA*	62	47.06%	0.01%
502	*MANDATORY*	62	50.00%	0.01%
503	*ENTAIL*	62	67.65%	0.01%
504	*DISTINCTION*	61	61.76%	0.01%
505	*CORE*	60	61.76%	0.01%
506	*CIRCULATION*	60	52.94%	0.01%
507	*SCHEDULE*	60	52.94%	0.01%
508	*ROMANIA*	59	41.18%	0.01%
509	*CAMPAIGN*	59	58.82%	0.01%
510	*VULNERABLE*	57	47.06%	0.01%
511	*WORLDWIDE*	57	58.82%	0.01%
512	*HIGHLIGHT*	57	67.65%	0.01%
513	*ALIGN*	57	44.12%	0.01%

Table 7 The least frequent fifteen word families of the EUWL

As can be seen in Table 8, the word families in the EUWL occurred in a wide range of the EU-related subject fields. Slightly more than 11% can be found in all the 34 subject fields and 299 (58.3%) word families occurred in 25 or more subject fields. Altogether 417 (81.3%) word families of the EUWL are found in 20 or more of the EU-related subject fields in the corpus under study.

Subject field covered	Number of word families	% of total EUWL
34	57	11.11%
33	31	6.04%
32	39	7.60%
31	23	4.48%
30	23	4.48%
29	22	4.29%
28	22	4.29%
27	28	5.46%
26	19	3.70%
25	35	6.82%
24	18	3.51%
23	23	4.48%
22	29	5.65%
21	28	5.46%
20	20	3.90%
19	14	2.73%
18	15	2.92%
17	20	3.90%
16	20	3.90%
15	9	1.75%
14	6	1.17%
13	7	1.36%
12	5	0.97%
Total	513	100.00%

Table 8 Subject field coverage of word families in EUWL

The detailed analysis of the word families of the EUWL found that these include legal words like *REGULATE* and *TREATY*, words in connection with funding like *FUND* and *RESOURCE*, the main EU institutions like *COMMISSION, PARLIAMENT* and *PRESIDENCY*. Besides the word list contains 15 (2.9%) abbreviations, for example, *DG, EC, OJ, SME* and 29 (5.7%) geographical names, which include all member states and names of two cities: *BRUSSELS* and *LISBON*, and a few – 5 (1%) – function words such as *PRIOR, BEHALF* and *VIA*.

3.2 Evaluation of the EUWL

The EUWL was evaluated for its specificity for EU discourse and relevance for English for EU purposes by testing its text coverage of texts representing several registers and genres. As it is shown in Table 9 the EUWL accounts for 18.03% of the tokens in the English EU Discourse Corpus. It is a fairly high coverage compared to earlier ESP word lists as the coverage of the AWL was reported to be 10.0% (Coxhead, 2000) and that of the MAWL was 12.24% (Wang et al., 2008). Besides the EEUD Corpus itself the EUWL was tested on four EU texts representing two different genres included in the corpus, namely, EU legislation and press releases. The four texts were published about a year later than other texts in the EEUD corpus and were selected at random. The EU word families account for 18.7% of the tokens of these EU texts (see: Table 10), which is very similar to the coverage of that of the original EEUD Corpus. In Appendix 2 a 500-word extract of an EU legislative text also illustrates the text coverage of the EUWL. The EUWL reached a high coverage – 17.02% – of another corpus of

EU texts, that was compiled according to different selection criteria than the EEUD corpus (Trebits, 2008). The EU English Corpus contains genres like information booklets, annual reports and sample test materials for recruitment competition for jobs in an EU institution, which are very different from the genres in the EEUD corpus. Thus the high coverage reinforces the validity of the EUWL as a word list useful for understanding EU texts in general.

Texts	Tokens	Text coverage	Mean text coverage
EEUD Corpus	1,076,460	18.03%	
EU English Corpus	197,620	17.02%	
EU corpora in total	1,274,080		17.52%
20th century US short story	4,575	1.44%	
20th century British play	7,873	0.17%	
20th century British novel	105,578	1.39%	
19th century British novel	1,013	2.37%	
Literary texts in total	119,039		1.34%

Table 9 Text coverage of EUWL in EU and literary texts

In order to establish that the EUWL is a truly EU-specific word list, it was also tested on literary texts, news texts, governmental and legislative texts. As can be seen in Table 9 on average elements of the EUWL accounted for 1.34% of four different literary genres of altogether almost 120,000 running words. Not surprisingly, this register seems to be the farthest from EU discourse. News texts with a slightly less than 5% coverage seem to apply a markedly different vocabulary than EU texts, which reinforces the findings of earlier research on contrasting the language in EU documents and news texts (Jablonkai, 2009).
In order to avoid the impact of text length on the results, the same analysis was carried out with the very first 500 tokens of the same texts. As can be seen in Table 10, these results do not show considerable differences in the tendencies revealed with whole texts.

Texts	Tokens	Coverage %	Coverage of first 500 tokens %
EU texts	5,326	18.70%	19.74%
News texts	3,567	5.33%	4.18%
UK government texts	2,017	13.29%	12.63%
UK legislation	13,497	19.25%	17.99%

Table 10 Text coverage of EUWL in different genres

As it is shown in Table 10, the EUWL, however, shows more than double the text coverage of governmental texts than news texts. An even higher text coverage was found in legislation texts. Not surprisingly, this shows a considerable similarity of EU discourse to these registers, especially, as regards their use of vocabulary. First of all, it can be explained by the similar formal, written style of these texts. Secondly, another reason is the contents of the EEUD Corpus, as EU legal texts account for about 63% of the whole corpus. However, taking a closer look at what vocabulary makes up this coverage, it turns out that the number of EU word families occurring in legislation and governmental texts is about 60% of the number of EU word families in EU texts (see: Table 11). Comparing tokens and number of EU word families statistically by the chi-square test shows that this difference is significant at $p=0.03$ level between EU texts and UK government texts and at $p=0.013$ level between EU texts and UK legislation texts. It suggests that although the text coverage of the EUWL in these texts is fairly high, they apply only a restricted group of EU word families.

Texts	Tokens	Coverage	Number of EU word families
EU texts	987	17.83%	75
UK government texts	917	12.63%	46
UK legislation	990	17.99%	47

Table 11 EUWL coverage and number of word families compared

Furthermore the word families of the EUWL were compared to word families in two other word lists. One of them was the AWL, which contains general academic vocabulary that is widely used in various disciplines (Coxhead, 2000). The other one is the BNC 3000 containing high frequency word families of the BNC. It is considered a general word list with some bias to written language based on the composition of the BNC, which contains 90% written and 10% spoken texts (Nation, 2004). The aim of the comparison was to test whether the word families of the EUWL can be considered EU-specific and whether it adds to the coverage of EU texts.

The comparison of the BNC 3000 and the EUWL showed that the total BNC 3000 contains about 60% of the word families in the EUWL. As can be seen in Table 12, most of the EU word families can be found among the second 1000 most frequent word families of the BNC. *AMEND, CLAUSE, COHERENCE, COOPERATE* and *REINFORCE* are examples of the word families that can be found in the EUWL but are not members of the BNC 3000. Contrasting the 513 word families of the EUWL to the 570 word families of the AWL showed that 323 word families overlapped. The words that can be found in both word lists include *COMMISSION, COMMUNITY, FINANCE, REGULATE, IMPLEMENT, PROCEED, POLICY, REQUIRE, AUTHORITY, ESTABLISH*. It means that almost 40% of the EU word families, like *ACCESSION, ACQUIS, CROSS-BORDER, ENLARGEMENT* and *RAPPORTEUR,* can be considered EU-specific. These findings are very similar to that of Wang and colleagues (2008) on the comparison of the MAWL and the AWL. Consequently, they strengthen their argument for the necessity of the development of subject-specific word lists for different disciplines. The words in the EUWL are not specific for any one sub-field of the EU's activities, therefore it can serve as a list of general EU vocabulary for students of EU studies.

Word list	Overlap in number of word families	Overlap in % of all EU word families
AWL	323	63%
BNC 3000	298	58.1%
BNC 1st 1000	77	15%
BNC 2nd 1000	167	32.6%
BNC 3rd 1000	54	10.5%

Table 12 Comparison of EU word list to the AWL and BNC 3000

A further argument for the application of the EUWL in ESP teaching is the high coverage of EU texts it provides. As it is shown in Table 13 the first 2000 word families of the GSL and the families of the EUWL together account for 93.5% of the EEUD corpus, that is already very close to the level of 95% coverage which is suggested necessary for understanding a text without a dictionary (Hirsch & Nation, 1992; Nation & Waring, 1997). Table 13 also shows that the EU-specific word families increase the text coverage by approximately 5% compared to the coverage of the general and academic word lists together.

Word lists	Coverage of EEUD Corpus
GSL+EUWL	93.56%
GSL+AWL	89.26%
BNC 3000	88.54%

Table 13 Text coverage of general and specific word lists

4 Pedagogical implications

It has been demonstrated that the EU word list comprises word families that are truly associated with English EU discourse, therefore it can serve as a reference for course and materials design for teaching English for the EU. At the same time, the EUWL provides guidelines for the sequencing of the teaching of lexical elements as it can follow the frequency order of word families in the list. With the help of the EU word list the EU-specific elements can be easily selected and can be used as the basis for traditional vocabulary teaching exercises and DDL activities focusing on the patterns specific to EU discourse.

It should be noted, however, that there are a few aspects of vocabulary that are not covered by word lists in general and by the EUWL in particular. One aspect is hyponymy, that is, a few lexical elements, for example *COUNCIL,* are not included in the EU word list, because they are elements of the GSL, although they have a specific meaning in the EU context. On the other hand, there are a few (11) elements which are included in the EU word list despite their general use, as they were not part of the GSL mainly because of its age. Examples here include *AUTOMATIC, WEBSITE, INTERNET*. The second aspect is multi-word units. Although many lexical items are used as part of bigger lexical units the concept of word list of word families, as applied in studies in ESP, concentrates on single-word lexical units. Therefore the EUWL should be complemented by multi-word items associated with EU discourse.

5 Conclusions

The present study has reported on the compilation of the EU word list based on the analysis of the EEUD Corpus that has been constructed based on the findings of a needs analysis questionnaire survey among EU professionals. The research has identified 513 EU word families and findings indicate that it provides a fairly high coverage of EU texts and it is specific to EU-related topics, consequently it can serve as a firm basis for courses of English for EU purposes.

The study has also reinforced the findings of earlier corpus-based analyses of the language of different disciplines, as it has found a marked difference between the elements of the EUWL and that of other general and academic word lists. On the one hand, it suggests that there is a considerable specificity in the discourse of disciplinary communities that can be characterised by their use of certain lexical items. On the other hand, it highlights the importance and relevance of developing specialised word lists in ESP for language teaching purposes for different disciplines.

References

Bauer, L., & Nation, P. (1993). "Word families". *International Journal of Lexicography, 6*(4), 253-279.

Biber, D. (2006). *University language a corpus-based study of spoken and written registers*. Amsterdam Philadelphia: John Benjamins Publishing Company.

Bowker, L., & Pearson, J. (2002). *Working with specialized language*. London, New York: Routledge.

Chen, Q., & Ge, G. (2007). "A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs)". *English for Specific Purposes, 26*(4), 502-514.

Chung, T. M., & Nation, P. (2004). "Identifying technical vocabulary". *System, 32*, 251-263.

Coxhead, A. (2000). "A new academic word list". *TESOL Quarterly, 34*(2), 213-238.

Ghadessy, P. (1979). "Frequency counts, word lists, and materials preparation: A new approach". *English Teaching Forum. 17*, 24-27.

Heatley, A., Nation, P., & Coxhead, A. (2002). *Range and frequency programs software*

Hirsch, D., & Nation, P. (1992). "What vocabulary size is needed to read unsimplified texts for pleasure?" *Reading in a Foreign Language, 8*(2), 689-696.

Jablonkai, R. (2008). Defining EU discourse. Steps towards the English EU Discourse Corpus. Paper presented at the SEAS PhD Conference. Eötvös Lóránd University. Budapest.

Jablonkai, R. (in press). "In the light of: A corpus-based analysis of two EU-related registers". *WoPaLP, 3.*

McArthur, T. (2003). World English, Euro-English, Nordic English? *English Today, 19*(1), 54-58.

Mudraya, O. (2006). "Engineering English: A lexical frequency instructional model". *English for Specific Purposes, 25*, 235-356.

Nation, P. (2001). "Using small corpora to investigate learner needs". In M. Ghadessy, A. Henry & R. L. Roseberry (Eds.), *Small corpus studies and ELT theory and practice*. Amsterdam Philadelphia: John Benjamins Publishing Company, 31-45.

Nation, P. (2004). "A study of the most frequent word families in the British National Corpus". In P. Bogaards, & B. Laufer (Eds.), *Vocabulary in a second language*. Amsterdam Philadelphia: John Benjamins Publishing Company, 3-13.

Nation, P. (2006). "How large a vocabulary is needed for reading and listening?". *The Canadian Modern Language Review, 63*(1), 59-82.

Nation, P., & Waring, R. (1997). "Vocabulary size, text coverage and word lists". In N. Schmitt, & M. McCarthy (Eds.), *Vocabulary description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 6-19.

Scott, M. (2004). *Oxford Wordsmith Tools version 4.0*. Oxford: Oxford University Press.

Sutarsyah, C., Nation, P. & Kennedy, G. (1994). "How useful is EAP vocabulary for ESP? A corpus-based study". *RELC Journal, 25*(2), 34-50.

Trebits, A. (2008). "English lexis in the documents of the European Union - A corpus-based exploratory study". *WoPaLP, 2*, 38-54.
Available at: http://langped.elte.hu/Wopalpindex.htm (accessed: 1 September 2009)

Trebits, A. (2009a). "Conjunctive cohesion in English language EU documents – A corpus-based analysis and its implications". *English for Specific Purposes, 28*(2), 199-210.

Trebits, A. (2009b). "The most frequent phrasal verbs in English language EU documents – A corpus-based analysis and its implications". *System, 37*(3), 470-481.

Wang, J., Liang, S., & Ge, G. (2008). "Establishment of a medical academic word list". *English for Specific Purposes, 27*(4), 442-458.

Ward, J. (1999). “How large a vocabulary do EAP Engineering students need?”. *Reading in a Foreign Language*, 12(2), 309-323.
West, M. (1953). *A General Service List of English words*. London: Longman.
Xue, G., & Nation, P. (1984). “A university word list”. *Language Learning and Communication. 3*(2), 215-229.

Appendices

Appendix 1

The one hundred most frequent word families in the EUWL

N	Headword	Range	Range in %	Cumulative family frequency	% of whole corpus
1	*EUROPE*	34	100,00%	7401	0,69%
2	*COMMISSION*	34	100,00%	5390	0,50%
3	*COMMUNITY*	34	100,00%	3635	0,34%
4	*REGULATION*	33	97,06%	2693	0,25%
5	*FINANCE*	34	100,00%	2693	0,25%
6	*IMPLEMENT*	34	100,00%	2285	0,21%
7	*PROCEED*	34	100,00%	2229	0,21%
8	*EC*	34	100,00%	2172	0,20%
9	*TREATY*	34	100,00%	1996	0,19%
10	*POLICY*	34	100,00%	1925	0,18%
11	*EU*	32	94,12%	1883	0,17%
12	*REQUIRE*	34	100,00%	1869	0,17%
13	*AUTHORITY*	34	100,00%	1864	0,17%
14	*ESTABLISH*	34	100,00%	1766	0,16%
15	*DIRECTIVE*	29	85,29%	1753	0,16%
16	*ENSURE*	34	100,00%	1645	0,15%
17	*PARAGRAPH*	32	94,12%	1611	0,15%
18	*OBJECTIVE*	34	100,00%	1595	0,15%
19	*COOPERATE*	34	100,00%	1586	0,15%
20	*PARLIAMENT*	34	100,00%	1576	0,15%
21	*PROJECT*	31	91,18%	1529	0,14%
22	*ECONOMY*	34	100,00%	1405	0,13%
23	*LEGISLATION*	34	100,00%	1320	0,12%
24	*FUND*	30	88,24%	1310	0,12%
25	*INSTITUTE*	33	97,06%	1283	0,12%
26	*AUDIT*	27	79,41%	1270	0,12%
27	*ENERGY*	26	76,47%	1259	0,12%
28	*CONTRIBUTE*	34	100,00%	1246	0,12%
29	*PROMOTE*	33	97,06%	1243	0,12%
30	*AREA*	34	100,00%	1227	0,11%
31	*ANNEX*	31	91,18%	1213	0,11%
32	*SPECIFIC*	34	100,00%	1203	0,11%
33	*LEGAL*	34	100,00%	1202	0,11%
34	*ENVIRONMENT*	31	91,18%	1171	0,11%
35	*REGION*	32	94,12%	1153	0,11%
36	*PERIOD*	34	100,00%	1101	0,10%
37	*APPROPRIATE*	34	100,00%	1098	0,10%
38	*BENEFIT*	32	94,12%	1096	0,10%
39	*FRAMEWORK*	34	100,00%	1043	0,10%
40	*DOCUMENT*	32	94,12%	1007	0,09%
41	*ISSUE*	34	100,00%	992	0,09%
42	*AMEND*	32	94,12%	989	0,09%

43	*SECURE*	33	97,06%	984	0,09%
44	*RELEVANT*	34	100,00%	967	0,09%
45	*STRATEGY*	31	91,18%	966	0,09%
46	*BUDGET*	34	100,00%	954	0,09%
47	*DATA*	33	97,06%	938	0,09%
48	*CONTRACT*	32	94,12%	923	0,09%
49	*SECTOR*	33	97,06%	909	0,08%
50	*OJ*	32	94,12%	874	0,08%
51	*GRANT*	32	94,12%	859	0,08%
52	*ASSESS*	34	100,00%	857	0,08%
53	*PRINCIPLE*	34	100,00%	853	0,08%
54	*COMMUNICATE*	34	100,00%	826	0,08%
55	*RESEARCH*	28	82,35%	817	0,08%
56	*PARTNER*	29	85,29%	810	0,08%
57	*ACCESS*	34	100,00%	805	0,07%
58	*PRIORITY*	32	94,12%	796	0,07%
59	*PROCESS*	34	100,00%	791	0,07%
60	*CULTURE*	27	79,41%	784	0,07%
61	*INVEST*	30	88,24%	781	0,07%
62	*FUNCTION*	32	94,12%	778	0,07%
63	*IDENTIFY*	34	100,00%	744	0,07%
64	*COORDINATE*	33	97,06%	742	0,07%
65	*SUBMIT*	34	100,00%	734	0,07%
66	*AVAILABLE*	33	97,06%	731	0,07%
67	*PARTICIPATE*	31	91,18%	729	0,07%
68	*ACHIEVE*	34	100,00%	720	0,07%
69	*ASSIST*	33	97,06%	710	0,07%
70	*CONSULT*	34	100,00%	700	0,07%
71	*AID*	27	79,41%	699	0,06%
72	*EVALUATE*	33	97,06%	697	0,06%
73	*DEFINE*	34	100,00%	694	0,06%
74	*INVOLVE*	34	100,00%	667	0,06%
75	*INITIATE*	34	100,00%	661	0,06%
76	*PURSUE*	32	94,12%	660	0,06%
77	*UNDERTAKE*	33	97,06%	658	0,06%
78	*CREATE*	34	100,00%	657	0,06%
79	*FINAL*	34	100,00%	656	0,06%
80	*NETWORK*	31	91,18%	640	0,06%
81	*INDICATE*	34	100,00%	639	0,06%
82	*MONITOR*	34	100,00%	636	0,06%
83	*RESOURCE*	33	97,06%	622	0,06%
84	*EXPENDITURE*	28	82,35%	616	0,06%
85	*INTEGRATE*	31	91,18%	611	0,06%
86	*MEDIUM*	29	85,29%	608	0,06%
87	*OBLIGATIONS*	32	94,12%	606	0,06%
88	*STRUCTURE*	33	97,06%	604	0,06%
89	*TECHNICAL*	34	100,00%	590	0,05%
90	*TRANSPORT*	30	88,24%	582	0,05%
91	*ELIGIBLE*	26	76,47%	575	0,05%
92	*PROTOCOL*	28	82,35%	570	0,05%

93	*ANNUAL*	32	94,12%	565	0,05%
94	*INTERNAL*	34	100,00%	552	0,05%
95	*REFERENCE*	34	100,00%	551	0,05%
96	*MAJOR*	33	97,06%	547	0,05%
97	*TECHNOLOGY*	30	88,24%	547	0,05%
98	*EEC*	27	79,41%	542	0,05%
99	*CHAPTER*	29	85,29%	539	0,05%
100	*TERRITORY*	31	91,18%	322	0,05%

Appendix 2

Text coverage of the EUWL in a 500-word extract of an EU legislation text
Elements of the EUWL are highlight in bold and are underlined.

COUNCIL **REGULATION (EC)** NO 284/2009
OF 7 APRIL 2009
AMENDING REGULATION (EC) NO 1083/2006 LAYING DOWN GENERAL PROVISIONS ON THE **EUROPEAN REGIONAL** DEVELOPMENT **FUND**, THE **EUROPEAN** SOCIAL **FUND** AND THE **COHESION FUND** CONCERNING CERTAIN PROVISIONS RELATING TO **FINANCIAL** MANAGEMENT
THE COUNCIL OF THE **EUROPEAN** UNION,
HAVING REGARD TO THE **TREATY ESTABLISHING** THE **EUROPEAN COMMUNITY**, AND IN PARTICULAR ARTICLE 161 THIRD **SUBPARAGRAPH THEREOF**,
HAVING REGARD TO THE PROPOSAL FROM THE **COMMISSION**,
HAVING REGARD TO THE ASSENT OF THE **EUROPEAN PARLIAMENT**,
HAVING REGARD TO THE OPINION OF THE **EUROPEAN ECONOMIC** AND SOCIAL COMMITTEE,
HAVING REGARD TO THE OPINION OF THE COMMITTEE OF THE **REGIONS**,
WHEREAS:

(1) THE UNPRECEDENTED **CRISIS** HITTING INTERNATIONAL **FINANCIAL** MARKETS HAS BROUGHT ABOUT **MAJOR CHALLENGES** FOR THE **COMMUNITY**, WHICH NECESSITATES A RAPID **RESPONSE** IN ORDER TO COUNTER EFFECTS ON THE **ECONOMY** AS A WHOLE AND, IN PARTICULAR, TO SUPPORT **INVESTMENTS** IN ORDER TO **PROMOTE** GROWTH AND EMPLOYMENT.

(2) THE **REGULATORY FRAMEWORK** FOR THE 2007-2013 PROGRAMMING **PERIOD** HAS BEEN ADOPTED WITH A VIEW TO **ACHIEVING** FURTHER SIMPLIFICATION IN THE PROGRAMMING AND MANAGEMENT OF THE **EUROPEAN REGIONAL** DEVELOPMENT **FUND**, THE **EUROPEAN** SOCIAL **FUND** AND THE **COHESION FUND**, THEIR EFFECTIVENESS AND **SUBSIDIARITY** IN TERMS OF THEIR **IMPLEMENTATION**.

(3) THE **ADAPTATION** OF CERTAIN PROVISIONS OF **REGULATION (EC)** NO 1083/2006 [1] IS NECESSARY IN ORDER TO **FACILITATE** THE MOBILISATION OF **COMMUNITY FINANCIAL RESOURCES** FOR THE START-UP OF OPERATIONAL PROGRAMMES AND **ASSISTED PROJECTS** WITHIN THE **FRAMEWORK** OF THESE PROGRAMMES IN SUCH A MANNER AS TO ACCELERATE **IMPLEMENTATION** AND THE **IMPACT** OF SUCH **INVESTMENTS** ON THE **ECONOMY**.

(4) IT IS NECESSARY TO STRENGTHEN THE POSSIBILITY OF PROVISION BY THE **EUROPEAN INVESTMENT** BANK (EIB) AND THE **EUROPEAN INVESTMENT FUND** (EIF) OF **ASSISTANCE** TO MEMBER STATES IN THE PREPARATION AND **IMPLEMENTATION** OF OPERATIONAL PROGRAMMES.

(5) TAKING ACCOUNT OF THE **STATUS** OF THE EIB AND EIF AS **FINANCIAL ENTITIES** RECOGNISED BY THE **TREATY**, WHEN **FINANCIAL** ENGINEERING OPERATIONS ARE ORGANISED **INVOLVING** THEM AS HOLDING **FUNDS**, IT SHOULD BE POSSIBLE TO DIRECTLY **AWARD** THEM A **CONTRACT**.

(6) IN ORDER TO **FACILITATE** THE USE OF **FINANCIAL** ENGINEERING INSTRUMENTS, NOTABLY WITHIN THE FIELD OF **SUSTAINABLE** URBAN DEVELOPMENT, IT IS NECESSARY TO PROVIDE FOR THE POSSIBILITY OF IN-KIND **CONTRIBUTIONS** BEING CONSIDERED AS **ELIGIBLE EXPENDITURE** IN THE **CONSTITUTION** OF, OR **CONTRIBUTIONS** TO, **FUNDS**.

(7) IN ORDER TO SUPPORT **ENTERPRISES**, AND IN PARTICULAR SMALL AND MEDIUM-SIZED **ENTERPRISES**, IT IS ALSO NECESSARY TO MAKE MORE **FLEXIBLE** THE CONDITIONS

GOVERNING THE PAYMENT OF ADVANCES WITHIN THE **FRAMEWORK** OF STATE AIDS UNDER ARTICLE 87 OF THE **TREATY**.

(8) IN ORDER TO ACCELERATE THE **IMPLEMENTATION** OF **MAJOR PROJECTS**, IT IS NECESSARY TO ALLOW **EXPENDITURES** RELATING TO **MAJOR PROJECTS** WHICH HAVE NOT YET BEEN ADOPTED BY THE **COMMISSION** TO BE INCLUDED IN **EXPENDITURE** DECLARATIONS.

(9) TO BOLSTER THE **FINANCIAL RESOURCES** OF MEMBER STATES THUS **FACILITATING** THE RAPID START-UP OF OPERATIONAL PROGRAMMES IN A **CRISIS CONTEXT**, IT IS NECESSARY TO **AMEND** THE PROVISIONS CONCERNING **PRE-FINANCING**.